ABSTRACT OF THE DISCLOSURE

A short latency and high bandwidth memory includes a systolic memory that is sub-divided into a plurality of memory arrays, including banks and pipelines that access these banks. Shorter latency and faster performance is achieved with this memory, because each bank is smaller in size and is accessed more rapidly. A high throughput rate is accomplished because of the pipelining. Memory is accessed at the pipeline frequency with the proposed read and write mechanism. Design complexity is reduced because each bank within the memory is the same and repeated. The memory array size is re-configured and organized to fit within desired size and area parameters.